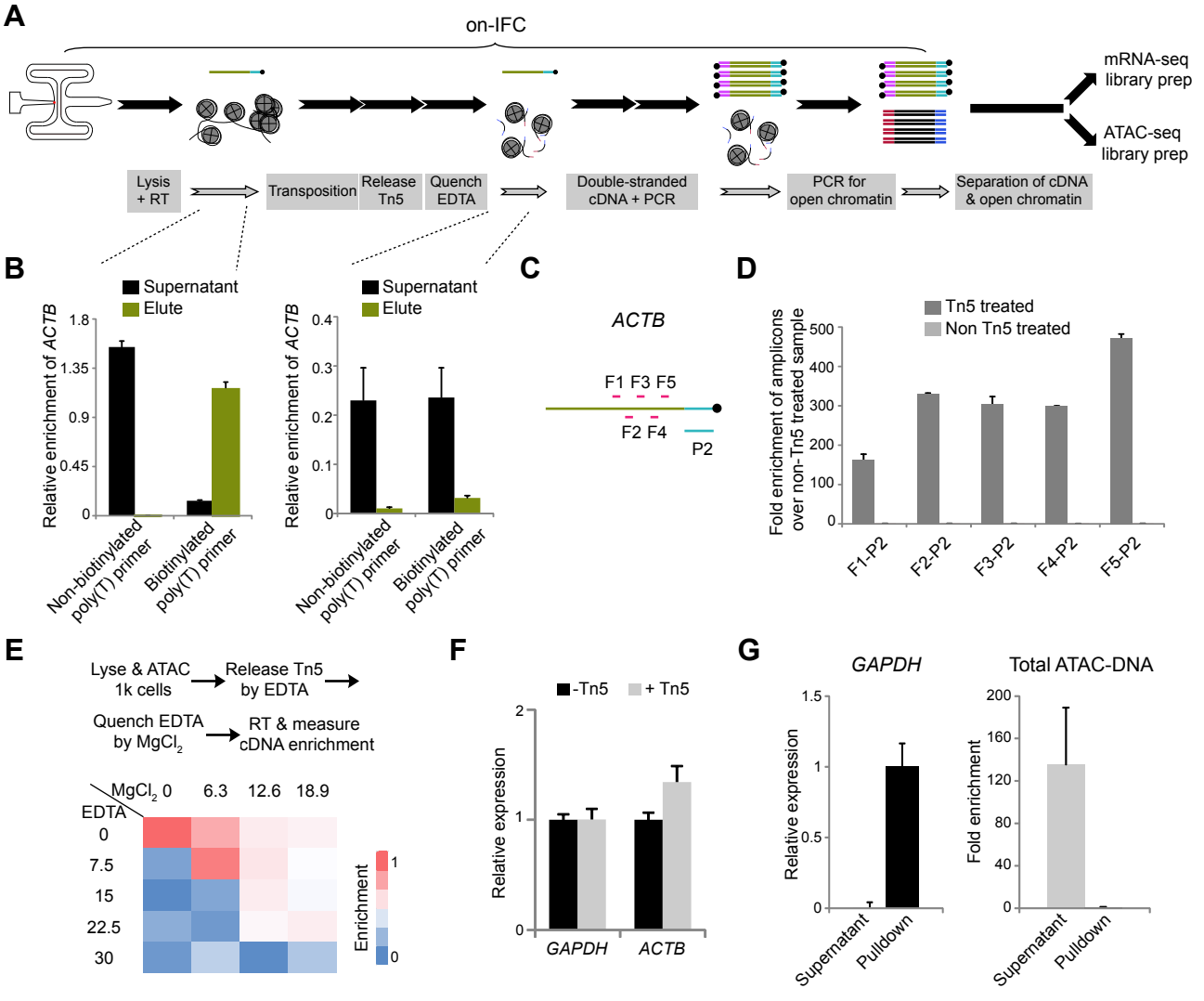
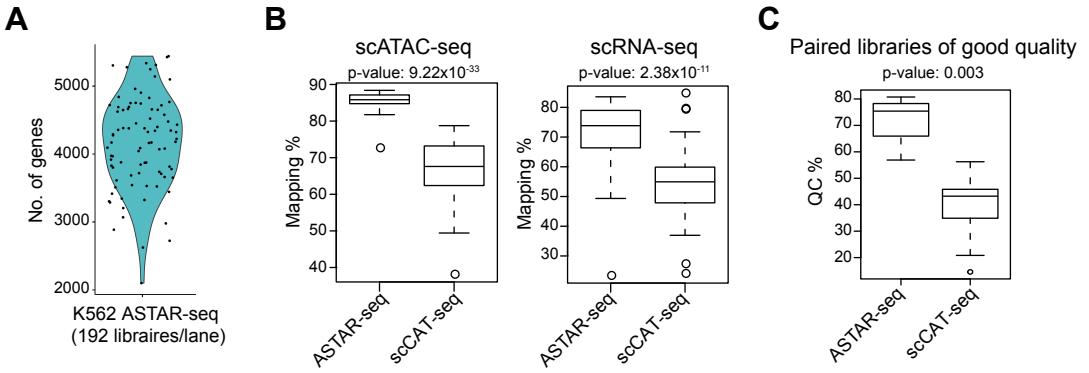


Supplemental Figure 1

Supplemental Figure 1. Earlier prototype and optimization of ASTAR-seq platform

(A) Overview of ASTAR-seq prototype. Briefly, cells were lysed and mRNA was reverse transcribed to single-stranded cDNA with biotin tag. Open chromatin was then tagmented by transposases Tn5 and inserted with sequencing adaptors, followed by inactivation of its enzymatic activity. Next, single-stranded cDNA was converted to double-stranded cDNA and amplified using biotinylated primers. Tagmented chromatin was then amplified using non-biotinylated primers targeting sequencing adaptors. Lastly, open chromatin and biotinylated cDNA were separated by streptavidin beads and processed for library preparation. (B) Left: Barchart showing relative enrichment of *ACTB* (Supplemental Table 5) in the supernatant and eluent of post RT samples separated by streptavidin beads. Biotinylated and non-biotinylated poly(T) primers were used for RT. Streptavidin beads pulled down most of the biotinylated cDNA, but not the non-biotinylated cDNA, suggesting the specificity of streptavidin beads. Error bar indicates SD, n=2. Right: Barchart showing relative enrichment of *ACTB* in the supernatant and eluent of samples post EDTA inactivation step separated by streptavidin beads. Majority of the biotinylated *ACTB* were not enriched in the eluent, an indication of Tn5 digesting the single-stranded cDNA. Error bar indicates SD, n=2. (C) Schematic of primers design for *ACTB* to affirm the digestion of single-stranded cDNA by Tn5. If single-stranded cDNA was digested by Tn5, digested fragments of cDNA would be further amplified by ATAC adaptors. (D) Barchart showing the relative enrichment of *ACTB* in the Tn5 treated sample over non-

Tn5 treated control, which were processed until “PCR for open chromatin” step following prototype ASTAR-seq protocol. *ACTB* primers F1-F5 (Supplemental Table 5) were used as the forward primer, and primer C1-P2-PCR (Supplemental Table 5) containing the same sequence as the poly(T) primer except for poly(T), was used as the reverse primer. Indeed, single-stranded cDNA was digested by Tn5, as fragments of cDNA were amplified by the ATAC adaptors. Error bar indicates SD, n=2. (E) Heatmap showing the enrichment of *ACTB* after inactivating Tn5 activity by different dosages of EDTA and quenching excess EDTA by variable amounts of $MgCl_2$. Schematic of the experimental design is shown on top. (F) Barchart showing relative enrichment of *GAPDH* (Supplemental Table 5) and *ACTB* in the samples processed following the pipeline specified in Fig.S1E, with or without addition of Tn5. cDNA amount in the Tn5 treated sample was comparable to the non-treated sample. Error bar indicates SD, n=2. (G) Barchart showing relative enrichment of *GAPDH* (left) and total amount of ATAC-DNA (right) in the supernatant and eluent of 1000 BJ cells processed following the pipeline in Fig.1A. Error bar indicates SD, n=2.



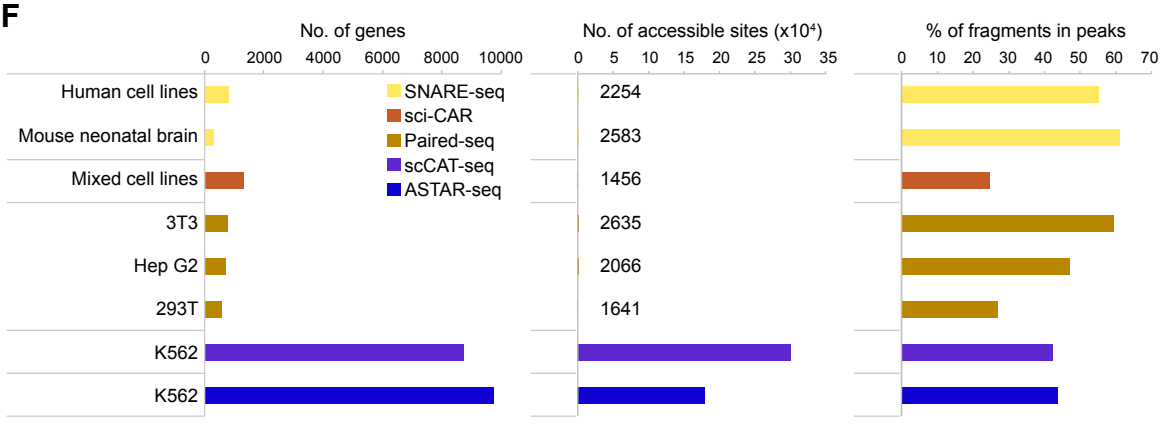
D

	SNARE-seq	sci-CAR	Paired-seq	scCAT-seq	ASTAR-seq
Total no. of cells profiled in the study	N.A.	N.A.	60,000	1,536	1,288
Total no. of libraries passing QC	16,433	16,121	30,646	621	935
Average % of libraries passing QC	N.A.	33%	49%	38.80%	72.60%

E

	SNARE-seq	sci-CAR	Paired-seq	scCAT-seq	ASTAR-seq
No. of cells	N.A.	2400	10000/sublibrary	96	96
No. of cells with good quality paired libraries	1000	800	3300	40	75
No. of sequencing lanes	2 (1 for mRNA; 1 for ATAC)	2 (1 for mRNA; 1 for ATAC)	2 (1 for mRNA; 1 for ATAC)	1	1
Chip cost (USD)	N.A.	N.A.	N.A.	N.A.	871
Reagents cost (USD)	653	3474	2210	2700	1407
Sequencing cost (USD)	2400	2400	2400	1200	1200
Overall cost (USD)	3053	5874	4610	3900	3478
Cost per good cell (USD)	3.1	7.3	1.4	97.5	46.4

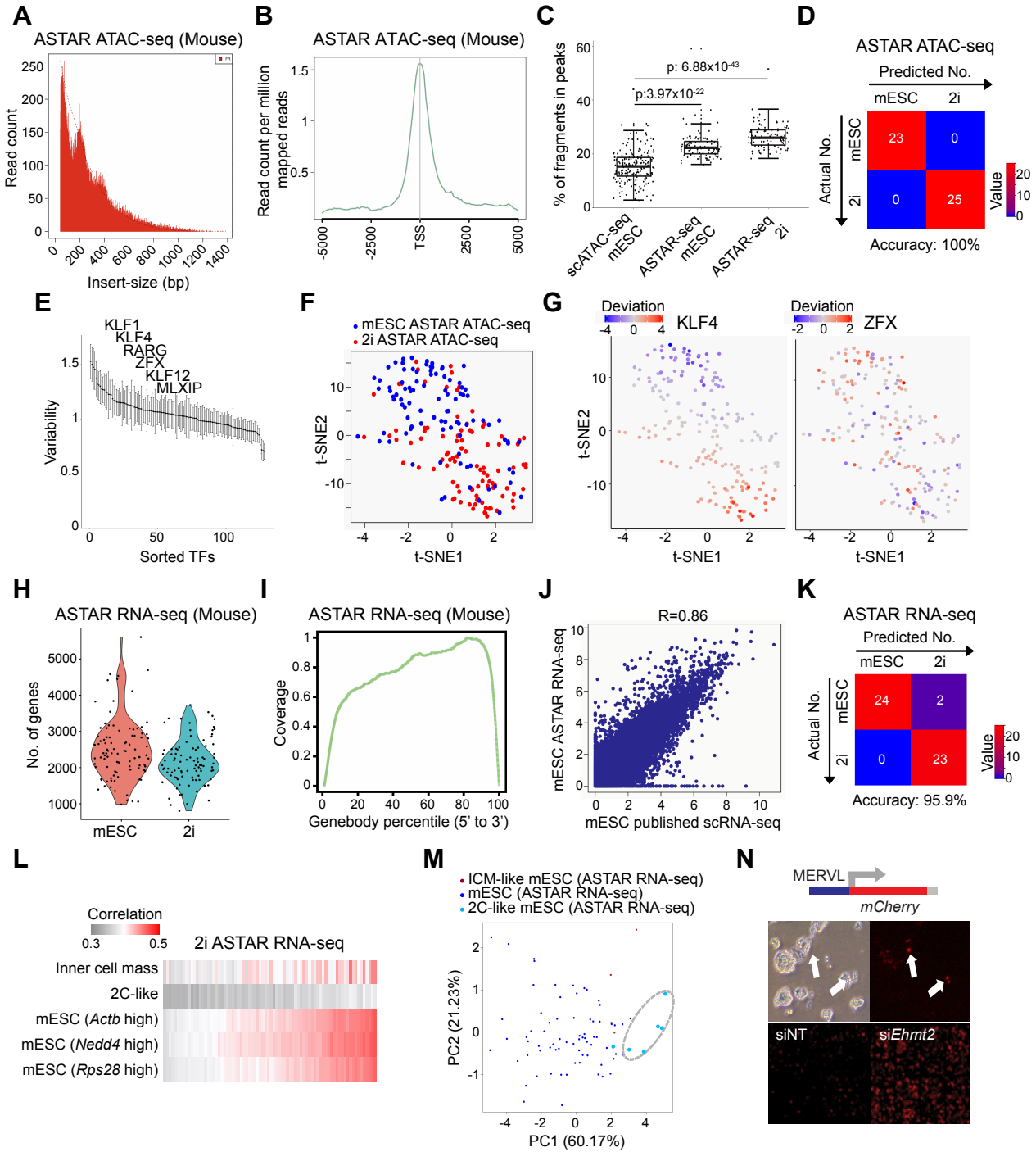
F



Supplemental Figure 2

Supplemental Figure 2. Comparison with other multi-modal single-cell techniques for chromatin accessibility and transcriptome

(A) Violin plot demonstrating the number of genes detected in K562 ASTAR RNA-seq libraries, which were sequenced at a depth of 192 single-cell libraries per lane of HiSeq 4000. (B) Boxplots showing the mapping percentage of K562 scATAC-seq (left) and scRNA-seq (right) libraries prepared following scCAT-seq and ASTAR-seq protocol. Two tailed student *t*-test was used to calculate p-values. (C) Boxplot demonstrating QC rate of scCAT-seq and ASTAR-seq with good quality paired libraries. Two tailed student *t*-test was used to calculate p-values. (D) Table summarizing the number of cells profiled, number of libraries of good quality, and average QC rate reported in the bimodal single-cell studies. (E) Table showing the estimated cost of bimodal single-cell assays for chromatin accessibility and transcriptome. (F) Barcharts indicating the median number of detected genes (RNA), number of accessible sites (ATAC), and % of fragments in peaks (ATAC) for single-cell libraries prepared by the respective bimodal techniques (Cao et al. 2018; Chen et al. 2019; Liu et al. 2019; Zhu et al. 2019).

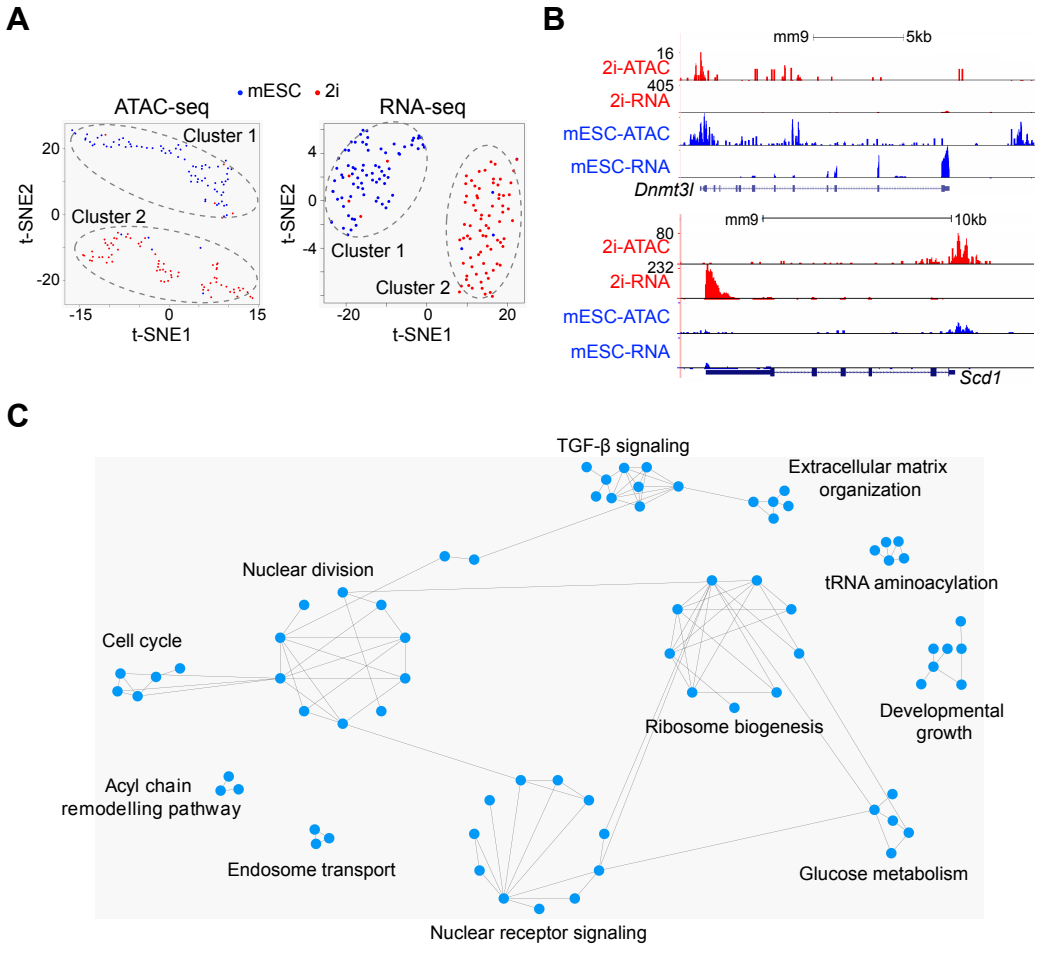


Supplemental Figure 3

Supplemental Figure 3. Heterogeneity of mESC and 2i cells in terms of chromatin accessibility and transcriptome

(A) Histogram demonstrating the frequency (y-axis) of fragments with the indicated insert size (x-axis). (B) Average enrichment profile indicating the read count per million mapped reads of a mouse ASTAR ATAC-seq library around all transcription start sites (TSS) of genome with a window of -5K to 5K. (C) Boxplot showing % of fragments in peaks of published mESC scATAC-seq and ASTAR-seq libraries of mESC and 2i cells. Two tailed student *t*-test was used to calculate p-values. (D) Confusion matrix for ASTAR ATAC-seq libraries of mESC and 2i cells using Support Vector Machines with Radial Basis Function Kernel algorithm. (E) Line plot showing variability levels (y-axis) of all mouse JASPAR motifs on the determined HARs of mESC and 2i ASTAR ATAC-seq libraries. (F) t-SNE clustering of mouse ASTAR ATAC-seq libraries. (G) Superimposition of deviation scores for KLF4 (left) and ZFX (right) motifs on the t-SNE plot. (H) Violin plot showing the number of genes detected in ASTAR RNA-seq libraries of mESC and 2i cells. (I) Line plot representing the coverage ratio (y-axis) of mouse ASTAR RNA-seq reads over the genebodies of housekeeping genes (x-axis). (J) Dotplot demonstrating Pearson's correlation between mESC ASTAR RNA-seq and published mESC scRNA-seq libraries. (K) Confusion matrix for ASTAR RNA-seq libraries of mESC and 2i cells using Random Forest algorithm. (L) Heatmap revealing correlation levels of each 2i ASTAR RNA-seq (x-axis) to various lineages of MCA panel (y-axis). Color indicates the correlation level, ranging from grey (low) to red (high). (M) PCA clustering of mESC ASTAR RNA-

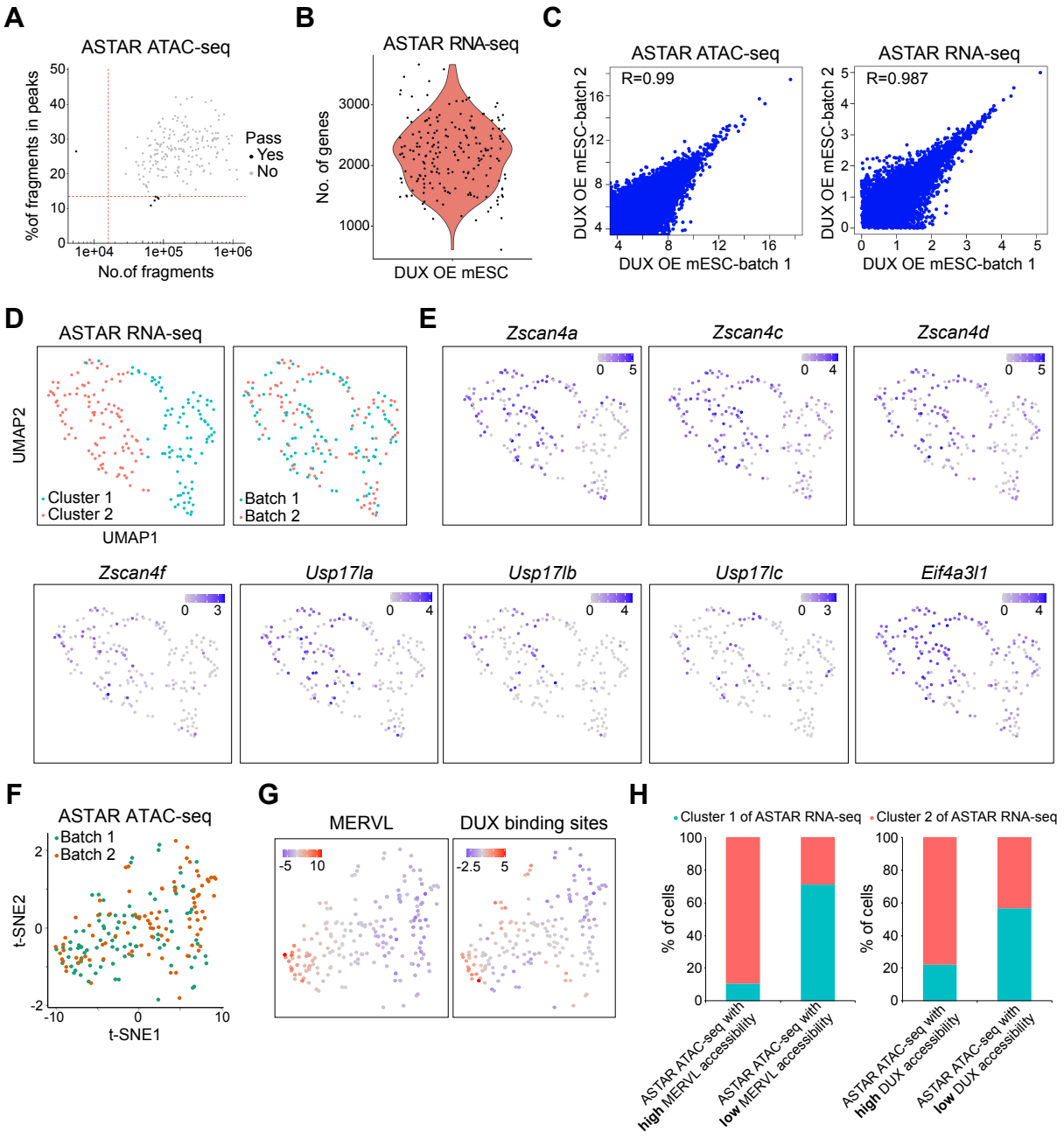
seq libraries based on MCA correlation values. Dotted ellipse surrounds the 2C-like mESCs. (N) Top: Schematic of the 2C reporter. Middle: Bright field (left) and fluorescence (right) images of mESCs transfected with 2C reporter. Bottom: Fluorescence images of mESCs with 2C reporter upon transfection with siNT control (left) and si*Ehmt2* (right). White arrows indicate the cells in which 2C reporter is activated.



Supplemental Figure 4

Supplemental Figure 4. Joint analysis of ASTAR-seq

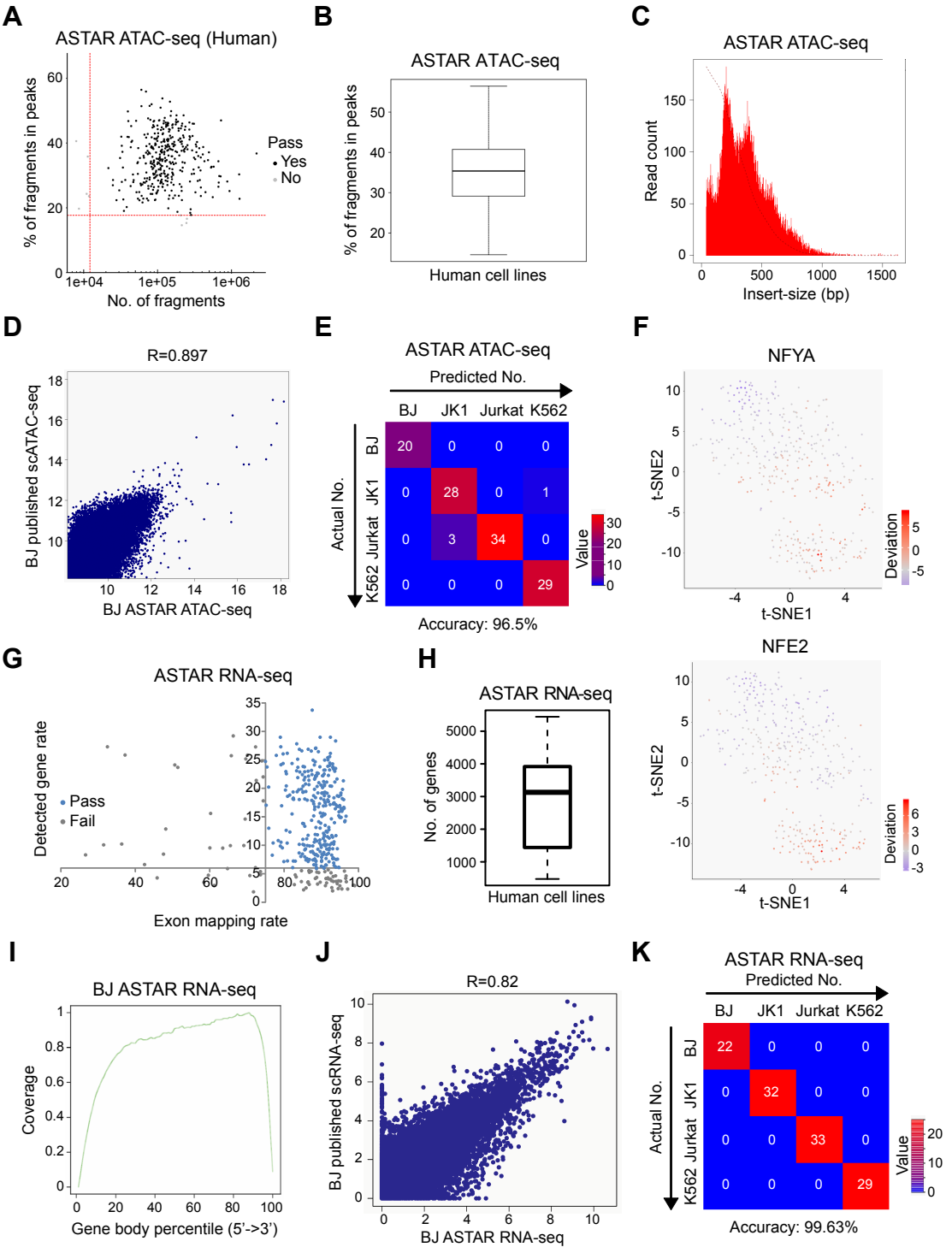
(A) t-SNE clustering of mESC and 2i ASTAR ATAC-seq libraries (left) and ASTAR RNA-seq (right) libraries, based on the differential accessible chromatin regions and differentially expressed genes identified by NMF analysis, respectively. (B) UCSC screenshots demonstrating the chromatin accessibility and expression levels of genes that are differentially accessible and expressed between the NMF clusters. (C) Interactome analysis showing the strong interaction among the cluster 1 specific genes through the specified pathways.



Supplemental Figure 5

Supplemental Figure 5. Transcriptome and chromatin accessibility landscape of 2C-like mESCs

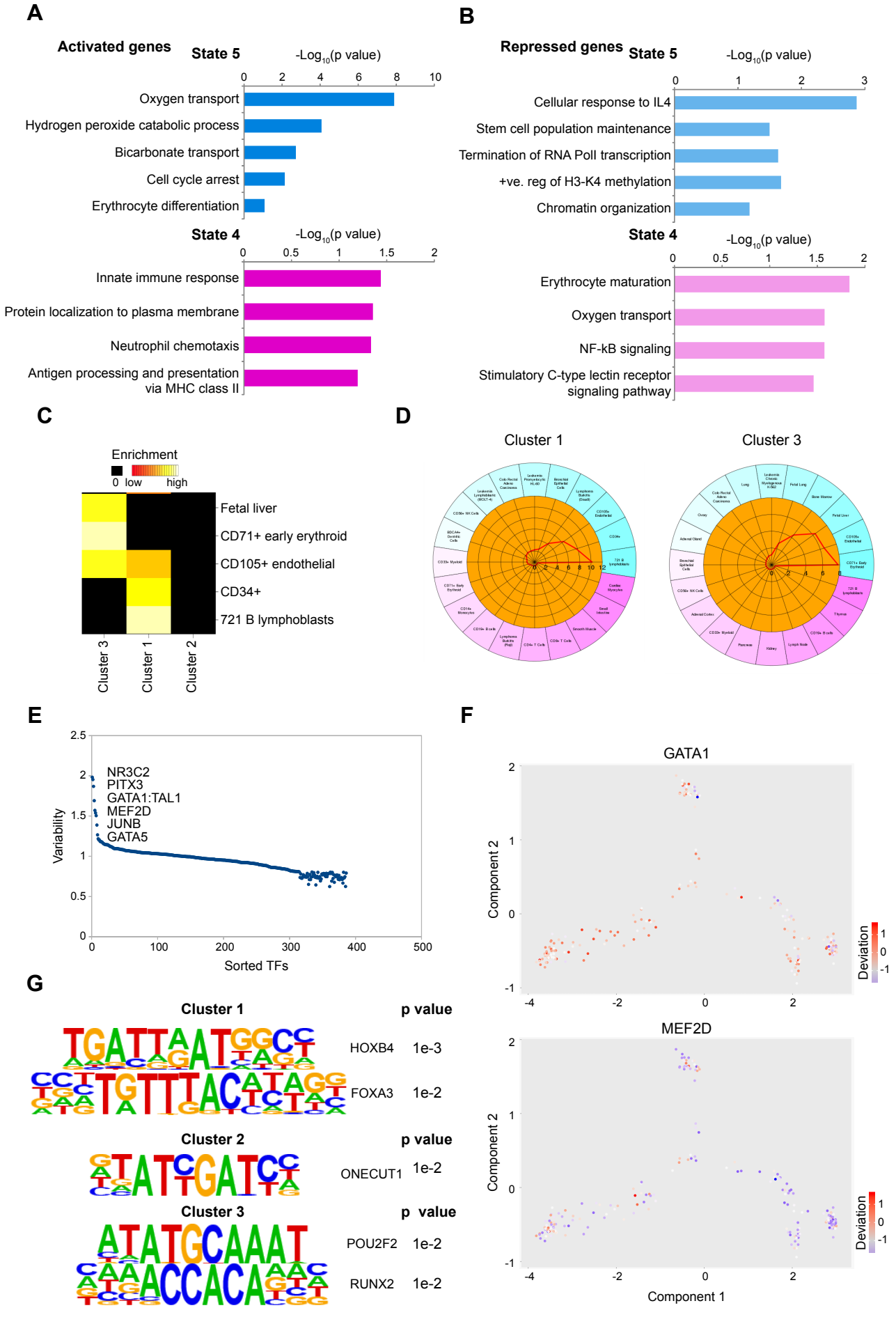
(A) Dotplot revealing % of fragments in peaks (y-axis) against the number of fragments (x-axis) of each ASTAR ATAC-seq library prepared from mESCs with DUX overexpression for 24hrs. Red dotted lines represent threshold values set for the respective criterion. (B) Violin plot showing the number of genes detected in ASTAR RNA-seq libraries of mESCs with DUX overexpression for 24hrs. (C) Dotplots demonstrating Pearson's correlation between the replicate ASTAR ATAC-seq (left) and ASTAR RNA-seq libraries (right) of mESCs with DUX overexpression for 24hrs. (D) UMAP clustering of ASTAR RNA-seq libraries prepared from mESCs with DUX overexpression for 24hrs. UMAP plots demonstrate the cluster (left) and batch ID (right) of each ASTAR RNA-seq library. (E) Superimposition of 2C-genes' expression on the UMAP plot. (F) t-SNE clustering of ASTAR ATAC-seq libraries prepared from mESCs with DUX overexpression for 24hrs. t-SNE plot indicates the batch ID of each ASTAR ATAC-seq library. (G) Accessibility levels of MERVL elements (left) and DUX binding sites (right) in the ASTAR ATAC-seq libraries of mESCs with DUX overexpression for 24hrs. (H) Annotation of ASTAR ATAC-seq libraries belonging to the indicated categories, based on the ASTAR RNA-seq clusters of the respective cells.



Supplemental Figure 6

Supplemental Figure 6. Quality control for ASTAR-seq libraries of human cell lines

(A) Dotplot revealing % of fragments in peaks (y-axis) of each human ASTAR ATAC-seq library plotted against the size of each library (x-axis). Red dotted lines represent thresholds set for the respective criterion. (B) Boxplot showing % of fragments in peaks of each human ASTAR ATAC-seq library. (C) Histogram demonstrating the frequency (y-axis) of fragments with the indicated insert size (x-axis). (D) Pearson's correlation between BJ ASTAR ATAC-seq and published BJ scATAC-seq. (E) Confusion matrix for ASTAR ATAC-seq libraries of human cell lines using Random Forest algorithm. (F) Superimposition of motif enrichment scores for NFYA and NFE2 on the t-SNE cluster in Fig.3d. Color indicates the enrichment level, ranging from blue (no) to red (high). (G) Dotplot revealing detected gene rate (y-axis) of each human ASTAR RNA-seq library plotted against exon mapping rate (x-axis). Blue dots represent the libraries which pass the QC, whereas grey dots represent the libraries that are excluded from downstream analysis. (H) Boxplot showing the number of genes detected in ASTAR RNA-seq libraries of human cell lines. (I) Line plot representing the coverage ratio (y-axis) of ASTAR RNA-seq libraries of human cell lines over the genebodies of housekeeping genes (x-axis). (J) Pearson's correlation between BJ ASTAR RNA-seq and published BJ scRNA-seq. (K) Confusion matrix for ASTAR RNA-seq libraries of human cell lines using Random Forest algorithm.



Supplemental Figure 7

Supplemental Figure 7. Characterization of cells belonging to each NMF cluster

(A-B) Barcharts showing the top GO terms enriched by the genes activated (A) or repressed (B) in cells of state 5 (top) or state 4 (bottom) as compared to the cells of state 1. (C) Heatmap demonstrating the lineages enriched by the NMF cluster specific genes using CTen. Color represents the enrichment score, ranging from black (no) to red (low) and to yellow (high). (D) Pie charts showing the lineages enriched by the NMF cluster 1 (left) and cluster 3 (right) genes. (E) Dot plot showing the variability levels (y-axis) of all human JASPAR motifs on the HARs determined from ASTAR ATAC-seq libraries of cells undergoing erythroblast differentiation. (F) Superimposition of motif enrichment scores for GATA1 and MEF2D on the trajectory plot for erythroblast differentiation. Color represents the enrichment level, ranging from blue (no) to red (high). (G) TF motifs enriched on cluster 1 (top), cluster 2 (middle), and cluster 3 (bottom) specific accessible regions. p values are indicated on the right.

SUPPLEMENTAL METHODS

Cell Culture Medium Recipes

mES-E14TG2a mouse embryonic stem cells were maintained in DME+4500 mg/l medium (HyClone) supplemented with 15% Embryonic Stem-cell FBS (Gibco), MEM Non-Essential Amino Acids Solution (100×, Gibco), 200mM L-Glutamine (100×, Gibco), 0.1mM β-mercaptoethanol (Gibco), 10000U/ml Penicillin-Streptomycin (100×, Gibco), 10^7 unit/ml ESGRO® Recombinant Mouse LIF Protein (10000×, Sigma-Aldrich).

2i cells were induced from mES-E14TG2a cells with N2B27 based 2i medium, which were harvested at passage 3 for ASTAR-seq. Medium recipe is as follows: DMEM/F-12 medium (Gibco), Neurobasal™ medium (Gibco), N-2 Supplement (200×, Gibco), B-27 Supplement (100×, Gibco), 200mM L-Glutamine (200×, Gibco), 0.1mM β-mercaptoethanol (Gibco), 7.5% BSA (1500×, HyClone), 3μM CHIR99021 (STEMCELL Technologies), 1μM PD0325901 (STEMCELL Technologies), 10^7 unit/ml ESGRO® Recombinant Mouse LIF Protein (10000×, Sigma-Aldrich).

mES-E14TG2a carrying inducible DUX overexpression construct was maintained using mESC medium with the addition of puromycin (1ug/ml) and G418 (500ug/ml). DUX overexpression was induced for 24hrs by adding dox (1ug/ml), which were then harvested for ASTAR-seq.

Human neonatal fibroblast cell line BJ (Stemgent, Cambridge, MA), were maintained in BJ medium, which is composed of DME+4500 mg/l medium (HyClone) supplemented with 10% Fetal Bovine Serum (Gibco), MEM Non-Essential Amino Acids Solution (100×, Gibco), 200mM L-Glutamine (100×, Gibco), 10000U/ml Penicillin-Streptomycin (100×, Gibco).

K562 (ATCC® CCL-243™) were maintained in K562 medium, which is composed of RPMI-1640 Medium with L-glutamine (HyClone) supplemented with 10% Fetal Bovine Serum (Gibco), 200mM L-glutamine (100×, Gibco), 10000 U/ml Penicillin-Streptomycin (100×, Gibco).

Jurkat, Clone E6-1 (ATCC® TIB-152™) were maintained in RPMI-1640 Medium with L-glutamine (HyClone) supplemented with 10% Fetal Bovine Serum (Gibco).

JK-1 (ACC 347) were maintained in RPMI-1640 Medium with L-glutamine (HyClone) supplemented with 20% Fetal Bovine Serum (Gibco).

Isolation of Mononuclear Cells and Erythroblast Differentiation

Umbilical cord blood samples were collected from Singapore Cord Blood Bank (SCBB; NUS IRB B-15-051), from which mononuclear cells were isolated by density gradient centrifugation using Ficoll® Paque Plus following the manufacturer guide (GE Healthcare, Uppsala, Sweden).

Mononuclear cells were cultured in Serum Free Expansion Medium II (SFEM II, STEMCELL Technologies, Vancouver, Canada) with Recombinant Human IL-3 (10ng/ml; PeproTech, NJ, USA), Recombinant Human IL-6 (10ng/ml; PeproTech, NJ, USA), Recombinant Human SCF (50ng/ml; PeproTech, NJ, USA), Recombinant Human Erythropoietin/EPO (2U/ml; R&D Systems, MN, USA), Recombinant Human IGF-I (40ng/ml; PeproTech, NJ, USA), Dexamethasone (1mM; Sigma-Aldrich, MO, USA), and L-Ascorbic acid (50ug/ml; Sigma-Aldrich, MO, USA). Media was replenished every other day for a period of 14 days.

Dead cells were excluded from the primary cells undergoing erythroblast differentiation following the manual of Dead Cell Removal Kit (Miltenyi Biotec). Live cells were collected for ASTAR-seq library preparation.

ASTAR-seq (Part 1: On-IFC Procedures)

We used C1 Single-cell Auto Prep System with Open App™ program (Fluidigm) and developed a novel ASTAR-seq protocol to prepare ATAC-seq and mRNA-seq libraries within the same single-cell.

Firstly, single-cell suspensions were prepared, washed once with C1 Cell Wash Buffer (Fluidigm), and diluted to a concentration within the range of 300-600 cells/µl, which was then mixed with C1 Cell Suspension Reagent (Fluidigm) at a ratio of 3:2 to prepare cell mix. Then, 20µl cell mix was loaded onto Fluidigm IFC for single-cell capture. Single cells were captured on C1 Single-cell Auto Prep Open App IFC using 'ASTAR- Cell Load (1861x/1862x/1863x)' script, which was generated using C1™ Script Builder software. Optionally, 'ASTAR- Cell Load and Stain (1861x/1862x/1863x)' script can be used, if LIVE/DEAD viability dyes (1:1000 Calcein-AM and 1:1000 ethidium homodimer-1 (Thermo Fisher Scientific) diluted with C1 Cell Wash Buffer (Fluidigm)) were added onto IFC for cell viability assessment.

Once the script run was completed, all cell-capture sites of IFC were imaged using Nikon automated microscope to measure the capturing efficacy.

Meanwhile, ASTAR-seq reagent mixtures (lysis/ATAC mix, EDTA mix, $MgCl_2$ and RT mix, cDNA-PCR mix) were freshly prepared and loaded onto the designated

wells of IFC, according to the loading map of 'ASTAR- ASTAR (1861x/1862x/1863x)' script. Script for this step takes approximately 6 hours.

On IFC, lysis and transposition reaction was first performed at 37°C for 30 min, followed by inactivation of Tn5 by EDTA at 50 °C for 30 min and priming of poly(T) primers at 72 °C for 3 min, 4°C for 10 min, and 25°C for 1 min. Next, RT was carried out at 50°C for 60 min, followed by heat inactivation of SSIV reverse transcriptase at 80°C for 10 min. In RT mix, $MgCl_2$ was added to quench excess EDTA from the previous inactivation step. Then, double-stranded cDNA was amplified using the following conditions: 98°C for 3 min; 5 cycles at 98°C for 20s, 58°C for 4min, and 68°C for 6min; and a final extension step at 72°C for 10 min. In cDNA-PCR mix, Bio-C1-P2-PCR-2 primer was used to biotinylate cDNA during PCR, which allowed for segregation of cDNA from ATAC-DNA using streptavidin beads. ATAC-DNA and amplified cDNA of each cell were harvested with 3.5µl (approximately) C1 Harvest Reagent (Fluidigm) from IFC and transferred onto a 96-well PCR plate.

Recipe of reagent mixtures are as follows:

(1) lysis/ATAC mix: 0.15% NP40 (Sigma-Aldrich), 1.5×Tagment DNA Buffer (Nextera DNA library preparation kit, Illumina), 1.5×Nextera® Tagment DNA enzyme 1 (Nextera DNA library preparation kit, Illumina), 1.5×C1 No Salt Loading Reagent (Fluidigm), 1.5U RNasin@ Plus RNase Inhibitor (Promega).
(2) EDTA mix: 8.75mM dNTP mix (Thermo Fisher Scientific), 8.5µM Bio-C1-P2-T31 primer (Supplemental table 5; IDT), 1×C1 No Salt Loading Reagent (Fluidigm), 2.6U RNasin@ Plus RNase Inhibitor (Promega), 41.25mM DTT (Thermo Fisher Scientific), 18.75mM EDTA (1_{st} BASE).
(3) $MgCl_2$ and RT mix: 2.93×SSIV Buffer (Thermo Fisher Scientific), 1×C1 No Salt Loading Reagent (Fluidigm), 3.5U RNasin@ Plus RNase Inhibitor (Promega), 35U SuperScript™IV Reverse Transcriptase (Thermo Fisher Scientific), 7.995µM C1-P2-RNA-Tso primer (Supplemental table 5; IDT), 22.05mM $MgCl_2$ (Thermo Fisher Scientific).
(4) cDNA-PCR mix: 1.233×NEBNext® Ultra™ II Q5 PCR master mix (NEB), 1.042µM Bio-C1-P2-PCR-2 primer (Supplemental table 5; IDT), 1×C1 No Salt Loading Reagent (Fluidigm).

ASTAR-seq (Part 2: Off-IFC Procedures)

1). *Separation of ATAC-DNA and cDNA*

500µl of streptavidin magnetic beads (Dynabeads™ MyOne™ Streptavidin C1, Thermo Fisher Scientific) were washed twice with 2×Binding and Washing buffer

(2M NaCl, 10mM Tris-HCl (pH 7.5), 0.02% Tween-20) and re-suspended with 500µl 2×Binding and Washing buffer and 250µl C1 DNA Dilution Reagent (Fluidigm). Then, 7.5µl beads mix were mixed with the 3.5µl single-cell samples harvested in the 96-well plate and incubated at room temperature for 20 min on a rotator. 10µl of cleared supernatant (ATAC-DNA) were collected and transferred to a new 96-well PCR plate (ATAC-seq plate). Beads were washed twice with 100µl 1×Binding and Washing buffer (1M NaCl, 5mM Tris-HCl (pH 7.5), 0.01% Tween-20), each for 5min. Beads were washed with 100µl 1×TE buffer (10mM Tris-HCl (PH 7.5), 0.02% Tween-20), which were immediately removed. 10µl of nuclease-free water was added to each well and the 96-well plate was immediately processed for next step.

2). *Further Amplification of cDNA and Clean-up*

15µl of cDNA-PCR master mix (1×NEBNext® Ultra™ II Q5 PCR master mix (NEB), 1.6µM C1-P2-PCR2 primer) was added to each well of the above 96-well plate containing beads and cDNA. PCR program used is as follows: 98°C for 3 min; 9 cycles of 98°C for 20s, 64°C for 30s, and 68°C for 6 min; 11 cycles of 98°C for 30s, 64°C for 30s and 68°C for 7 min; and a final extension at 72°C for 10 min.

22.5µl AMPure XP magnetic beads (Beckman Coulter) were added to each well and incubated at room temperature for 10 min. AMPure XP beads were washed twice with freshly prepared 75% ethanol and eluted with 11µl nuclease free water. 10µl supernatant were then transferred to a new 96-well PCR plate.

3). *Preparation of ASTAR RNA-seq Libraries for High-throughput Sequencing*

To quantify average concentration of cDNA within the size range of 200-9000bp, amplified cDNA of 11 single-cells were measured using Agilent High Sensitivity DNA Kit (Agilent Technologies). cDNA was then diluted to a final concentration within the range of 0.15-0.2 ng/µl for mRNA-seq library preparation following "Using C1 to Generate Single-Cell cDNA Libraries for mRNA Sequencing" manual.

Briefly, 1.25µl of diluted cDNA was mixed with 2.5µl Tagment DNA buffer (Nextera XT DNA Library Prep Kit, Illumina) and 1.25µl Amplification Tagment Mix (Nextera XT DNA Library Prep Kit, Illumina) and incubated at 55°C for 10 min, followed by neutralization with 1.25µl NT Buffer (Nextera XT DNA Library Prep Kit, Illumina). 3.75µl NPM (Nextera XT DNA Library Prep Kit, Illumina), 1.25µl Nextera® XT index 1 (Nextera® XT Index Kit, Illumina), and 1.25µl Nextera® XT index 2 (Nextera® XT Index Kit, Illumina) were added to each well for barcoding cDNA of each cell. PCR program used is as follows: 72°C for 5min; 95°C for 30s; 12 cycles of 95°C for 10s, 55°C for 30s and 72°C for 1min; and a final extension at 72°C for 5 min. PCR products were pooled in a tube with a final volume of ~ 1.1ml.

Pooled ASTAR RNA-seq libraries were purified twice with AMPure XP magnetic beads (Beckman Coulter, 0.75×) with the same procedures as mentioned above. Quality of ASTAR RNA-seq libraries was assessed using Agilent DNA 7500 Kit (Agilent Technologies) and qPCR of genes of interest.

4). *Preparation of ASTAR ATAC-seq Libraries for High-throughput Sequencing*

10µl supernatant in the ATAC-seq plate were mixed with 90µl ATAC-PCR master mix (1×NEBNext® Ultra™ II Q5 PCR master mix (NEB), 1.14µM custom barcode adaptor 1 (Supplemental table 5, IDT), and 1.14µM custom barcode adaptor 2 (Supplemental table 5, IDT)). PCR program used is as follows: 72°C for 5min; 98°C for 30 s; 22 cycles of 98°C for 10 s, 72°C for 90 s; and a final extension at 72°C for 10 min.

PCR products were pooled at a final volume of ~ 9ml. ASTAR ATAC-seq libraries were precipitated by mixing with 32.2ml 100% ethanol and 4.6ml 3M sodium acetate, and incubating at -80°C overnight. The mixture was centrifuged at 15000g for 20 min at 4°C. DNA pellet was then washed with freshly prepared 75% ethanol twice and re-suspended with 200µl nuclease free H_2O. ASTAR ATAC-seq library was purified by MinElute PCR Purification Kit (QIAGEN) and eluted with 50µl nuclease-free H_2O. Next, ATAC-seq library was purified using AMPure XP magnetic beads (Beckman Coulter, 1.2×) and eluted with 30µl nuclease-free H_2O.

Quality of ASTAR ATAC-seq library was assessed using Agilent DNA 7500 Kit (Agilent Technologies).

Determination of Duplication Rates for ASTAR ATAC-seq and ASTAR RNA-seq libraries

The mapped files were used as inputs for the MarkDuplicates module of PICARD. The metrics files output of MarkDuplicates contained the information about the duplication percentage. Total reads count and mapped reads count for each library were obtained from the "log.final.out" output of STAR aligner.

Filtering scRNA-seq Libraries

BAM outputs were uploaded to SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). RNA-seq QC plot was generated. For mouse samples, libraries with gene detection rate below 15% and/or exon mapping percentage below 75 were filtered out. For human libraries, cells with gene detection rate below 6% were filtered out. If more than one read was detected in an exon of a gene, the gene was counted as a detected gene as per SeqMonk protocol.

Determination of DNA contamination in ASTAR RNA-seq Libraries

The fold difference between the enrichment of merged ASTAR RNA-seq and merged ASTAR ATAC-seq libraries over the determined HARs was measured using getDifferentialPeaks. Fold threshold was set to 3 using the –F option.

Gene Coverage

We merged the BAM files belonging to each category of cells using SAMtools merge (Li et al. 2009). Then we used the geneBody_coverage.py module of RSeQC (Wang et al. 2012) to determine the distribution of the mapped fragments over the genebodies of house-keeping genes.

UCSC Genome Browser Screenshots for scRNA-seq Libraries

The scRNA-seq libraries belonging to each cell type were merged using SAMtools merge. This was followed by the creation of tag directories using the "makeTagDirectory" script of HOMER (Heinz et al. 2010). Finally, the "makeUCSCfile" script was used and the options -style and -fragLength were set to rnaseq and given respectively.

Correlation with Mouse Cell Atlas (MCA)

FPKM table output of cuffnorm (mouse libraries) was uploaded to Mouse Cell Atlas (Han et al. 2018) (http://bis.zju.edu.cn/MCA/blast.html). The output obtained after the MCA analysis was used to create PCA graphs using FactoMineR (Lê et al. 2008) package in R (R Core Team 2019) and to categorize the populations of 2i and mESC ASTAR RNA-seq libraries. Threshold was set at 0.414 to consider a cell belonging to a particular lineage.

RCA Analysis

FPKM table output of cuffnorm (human libraries) was used as an input for RCA (Li et al. 2017). The human ASTAR RNA-seq libraries were clustered using the Global panel mode of RCA with default parameters. Detailed script is shown in Supplemental Codes - Bioinformatic scripts.

Gene Ontology Analysis

Lists of genes of interest were uploaded to DAVID (Dennis et al. 2003) (https://david.ncifcrf.gov/). The terms identified by the "GOTERM_BP_DIRECT" functional annotation were used to generate the bar charts.

Processing of scATAC-seq Libraries

Group information was added to each de-duplicated scATAC-seq library using the AddOrReplaceReadGroups module of PICARD. This was followed by lexicographical sorting of each library using ReorderSam module of PICARD. The option ALLOW_INCOMPLETE_DICT_CONCORDANCE was set to TRUE. The sequence dictionaries for hg19 and mm9 that were used for sorting lack ChrM and other ambiguous chromosomes. Each library was further indexed using SAMtools index.

Nucleosomal Pattern Determination

The histograms were generated using CollectInsertSizeMetrics module of PICARD.

Average Enrichment Profile

The average enrichment of ASTAR ATAC-seq libraries over TSS was determined using ngsplot (Shen et al. 2014).

Depth of Coverage Detection

Coverage of each processed library over the HARs was quantified using DepthOfCoverage module GATK tools v3.46 (McKenna et al. 2010). The option COUNT_FRAGMENTS was used. The values indicated in the total_cvg columns of the interval_summary outputs of GATK were used for downstream analysis.

For correlation with published scATAC-seq libraries, single cells belonging to each study were merged using SAMtools merge. Then the merged BAM files were subjected to further processing and filtering as mentioned above. The coverage values were then used to generate the correlation dotplots. The correlation values were calculated using EXCEL. Detailed script is shown in Supplemental Codes - Bioinformatic scripts.

chromVAR Analysis and t-SNE Clustering

The duplicates were removed from scATAC-seq libraries using the MarkDuplicates module of PICARD. Reads mapping to chromosome M and Y were removed. These libraries were then uploaded to chromVAR (Schep et al. 2017) along with the narrowPeaks file as an input for the getCounts function of chromVAR.

QC was performed using the filterSamples function. The cutoffs for scATAC-seq libraries were estimated based on the medians of each dataset. The minimum fragment of peaks threshold (“min_in_peaks”) was set to 0.5 times median % of

fragments in peaks. The minimum accepted depth of each library ("min_depth") was set to the maximum of 10% of median library size.

Motif variability over these HARs was measured using the computeDeviations function followed by the computeVariability option. scATAC-seq libraries were correlated using the getSampleCorrelation module. t-SNE clustering of scATAC-seq libraries was carried out using the deviationsTSNE option with a perplexity setting of 30. For meta-analysis, the published scATAC-seq libraries were processed similarly and included in the analysis.

For clustering scATAC-seq libraries based on MERVL and DUX binding sites, the computeDeviations was performed on these sites by using the function "getAnnotation". DUX binding sites were obtained from the previous study (Hendrickson et al. 2017) and converted to mm9 coordinates using the UCSC liftover tool. MERVL loci for mm9 were obtained from UCSC Table Browser.

Detailed script can be found in Supplemental Codes - Bioinformatic scripts.

Integrative Analysis and Associated Downstream Analysis

For CoupleNMF, K setting was determined by running the script starting with K=5 and then determining the best NMF score. We kept on reducing the K value and re-running the script until the NMF score obtained was > 1 (human: K=3, and mouse: K=2). The motifs enriched by the peaks specific to each NMF cluster were determined by findMotifsGenome.pl. Detailed script can be found in Supplemental Codes - Bioinformatic scripts.

For human ASTAR RNA-seq libraries, genes identified to be significantly expressed in each cluster were subjected to CTen (Shoemaker et al. 2012) (http://www.influenza-x.org/~jshoemaker/cten/) for identification of the enriched lineages. In addition, NMF clusters were superimposed on the pseudotime trajectory. On the other hand, mouse ASTAR RNA-seq libraries were clustered using Seurat (Butler et al. 2018) to ensure the correlation of ASTAR RNA-seq with NMF clusters.

For mouse ASTAR ATAC-seq, the cluster-specific regions were used to calculate deviations using the "getAnnotations" Function of chromVAR. ASTAR ATAC-seq heatmaps for NMF clusters were created by obtaining read counts over the NMF cluster-specific peaks using featureCounts (Liao et al. 2014) with -F SAF option. The values obtained were used to generate heatmaps using heatmap.2 function of gplots in R (R Core Team 2019).

Combined t-SNE (coupled NMF)

The genes and accessible regions that were considered significant by the NMF clustering were subsequently used in a combined matrix to cluster the cells based on both signals together using Seurat. Raw coverage counts for the accessible regions were generated using GATK (McKenna et al. 2010), as shown in Supplemental Codes - Bioinformatic scripts - Cicero.txt. The expression values of significant genes were added to the same matrix. The matrix was used as input for Seurat to cluster the cells.

Confusion Matrix Generation

The input for ATAC-seq was the raw depth of coverage values over HARs. HARs were determined and libraries were de-duplicated as mentioned above (Supplemental Codes - Bioinformatic scripts - Processing_ATAC_For_chromVAR.txt). Then GATK version 3.4-46 was used to measure the coverage of each cell over each HAR locus (McKenna et al. 2010), as shown in Supplemental Codes - Bioinformatic scripts - Cicero.txt. The input for RNA-seq was the raw counts table of each gene in each library. The raw counts table was generated using htseq count function (Anders et al. 2015). We partitioned the input files randomly into two parts. First part consisted of 70% of the input file and was used as training set. The other part (30% of the data) was used as a test set to determine the accuracy of ASTAR-seq in distinguishing among various cell types. The training data sets were classified using SVM, Random Forest and PLDA classifiers. The method that demonstrated the highest accuracy was used subsequently to classify the test sets and generate the confusion matrix using the MLSeq functions. Detailed script can be found in Supplemental Codes - Bioinformatic scripts.

Statement for Why GRCh38 and GRCm38 (mm10) Would Not Significantly Affect the Conclusions of This Study

1). The LiftOver tool of UCSC demonstrates that DUX binding sites and MERVL loci used for the analysis of DUX OE ASTAR-seq libraries can be lifted over from mm9 to mm10 with extreme success. The conversion succeeded in 37426 loci out of 37442 loci used in the manuscript (99.96% of the loci). The same can be said for all loci used in other ATAC-seq analyses throughout the manuscript. Hence, the conclusions based on ATAC-seq analyses will not be affected significantly by the choice of genome version.

2). All genes that can distinguish the cell lines used in the study can be found in both versions of the genome. Hence, the clustering outcomes will not be affected significantly by the choice of genome version and cell lines will be separated in a similar manner.

3). For comparison analyses (with previously published techniques), all libraries were mapped to the same genome index (mm9 for mouse and hg19 for human samples). Hence, the effect of genome version on the comparison analysis will be similar on all libraries and the conclusions will not be significantly affected.

REFERENCE

Anders S, Pyl PT, Huber W. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31:** 166–169.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4:** R60.

Lê S, Josse J, Husson F. 2008. FactoMineR: An R package for multivariate analysis. *J Stat Softw* **25:** 1.

Liao Y, Smyth GK, Shi W. 2014. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923-930.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

Shen L, Shao N, Liu X, Nestler E. 2014. Ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15:** 284.

Wang L, Wang S, Li W. 2012. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28:** 2184-2185.